Bilingual Translator : A System for Automatically Translating Utterances Mixed with Telugu & English

Final paper: NLP prototype design proposal, CSCI 544

Sudeeptha Mouni Ganji
USC ID: 2942771049
USC Email: sganji@usc.edu

November 22, 2020

Abstract

The advent of technology and migration of people across the world has resulted in people being multilinguists or having a basic knowledge of popular words or quotes. These words are used in social media in combination with the local language or as part of daily conversations. When such an utterance occurs, it is called code-switching if a whole clause is replaced and code-mixing if a few words are replaced. Often, these terms are used interchangeably. Such utterances create confusion and difficulty in reading for participants who know only one of the languages being used. While several sites implement translators, these translators do not work effectively on code-mixed and code-switched sentences. Telugu is the third most commonly spoken Indian language in the United States, but there are no translators for Telugu-English mixed utterances. This paper proposes a model to translate utterances in Telugu-English into a single language.

Keywords: Machine Translation, Code-switching, Code-mixing, Language Identification, NLP, Telugu-English translation

1. Introduction

In multilingual countries like India, many people speak more than one language and use a mixture of languages in different formal and informal settings. They try to adopt local or global languages by mixing in their own language resulting in a relatively stable mixture of two languages like Hinglish. When such people message or comment on apps, like Whatsapp, Instagram, Facebook, and Twitter, they switch between languages by adding quotes, words, or phrases. Consumers post reviews of products using a combination of local and global languages on international brands. While such utterances reflect local culture and preferences, it causes confusion to other users who know only one of the languages used in the sentences.

Social platforms that implement language translators, have translators suitable for only converting one language to another language but not for converting mixed languages. This is because the mixed language is difficult to translate as many foreign, transformed words are present in a single setting and the words should be identified, translated and made grammatically correct. As such utterances usually occur in informal settings, collecting sufficient data is a challenge. This paper suggests a model to translate everything into a single language when both Telugu and English are used in the same utterance.

2. Method

2.1 Materials:

Since, such conversations usually occur in informal settings, one way we could capture the data is by scraping the web and social media sites for sentences that use a mix of English and Telugu in a single utterance. Twitter API can be a useful source for this purpose as we can collect data and comments posted by bilingual Twitter users. This data will have to be cleaned and filtered so that only sentences containing English and Telugu together remain. Another option would be to use voluntary participants who have basic knowledge of English and Telugu to provide conversation snippets. They do not need to be fluent in both languages as they can use a combination of the languages to make up for a loss of appropriate words.

After getting the required data it can be observed that, the sentence format of Telugu-based sentences tends to be in the Subject-Object-Verb format while that of English-based sentences tends to be in the Subject-Verb-Object format. Additionally, the data can have different word replacements like

i) any of the verbs, adjectives, nouns or pronouns are in one language while remaining words are in another language,

ii) new words are formed by adding the plural form of one language to the stem word of another language i.e., Telugu word + English plural or English word + Telugu plural.

2.2 Procedure:

The procedure consists of three main steps.

i) Identifying the language of the words,

ii) Translating the words, and

iii) Modelling the output to make it more grammatically correct.

2.2.1 Identifying the language of the words

In this step, we would first pass the data through an analyzer which would annotate the English words by using an English corpus. We would then transliterate the remaining words to Telugu. These words would be passed through another analyzer to annotate the Telugu words by using a standard Telugu corpus. Once this is done, all the remaining unmarked words would be checked to see if they are a mix of stem words in Telugu and the plural forms in English or vice-versa. These words would be annotated accordingly.

Remaining words if any, can be considered as proper nouns if they start with a capital letter.

Another part in this step would be the identification of phrases. Since both Telugu and English support complex sentences, an utterance can have whole phrases in a single language. So, while executing the above steps we could also mark connective words in both languages. If all the words after a connective or comma or new sentence have a continuous set of words in a single language until the next connective, then it is highly likely that these set of words is a phrase. So, we can mark the whole phrases accordingly along with their languages.

If we know whether the data is code-mixed or code-switched, we can skip the first part or the second part of this step accordingly. But, if we don't know much about the data, we should perform both the steps.

2.2.2 Translating the words

Once the words and phrases have been marked, the code-mixed parts need to be translated. First translate all the phrases into Telugu. Then translate the Telugu stem-English plural words to Telugu stem plurals. For words with English stem-Telugu plural, first convert the English stem word to Telugu and then convert the word into plural form. Next, convert the English annotated words into Telugu. The words annotated as Telugu initially can remain the same. Now the whole sentence is converted to Telugu.

2.2.3 Modeling the output to make it more grammatically correct

Once the whole sentence is converted to Telugu, we can reorder the words by using a single language model. For our purpose, we can use a bigram or a trigram model which has been trained on a Telugu corpus which consists of correct Telugu sentences. The generated sentences are tokenized and then passed to the model for reordering and grammar correction. If the target language is Telugu, then this is the required output. If the target language is English, the sentences can then be converted to English by using a standard Telugu-to-English convertor.

2.3 Evaluation:

The results of our model could be compared with Google Translate to check the correctness of our sentences. Though Google Translate works upto an extent with mixed sentences, often different kinds of errors were observed upon my analysis. One error was upon using auto-detect language with the input of English-Telugu mixed sentences, different languages like Slovan, Japanese, Yoruba etc were detected and the translations were wrong. Another error was when the input language was specified as Telugu, there were still grammatical and translation errors.

Another option would be using human participants to check for correctness of utterances. The following point-based system can be used.

0	Sentences are not understandable
1	Sentences are understandable but have grammatical errors.
2	Grammatically correct, understandable and simple sentences.
3	Grammatically correct, understable, simple and complex sentences.
4	Grammatically correct, understandable, simple and complex sentences with appropriate connectives.

The average score obtained can be used as a measure for efficiency of our model. For more robust checking, users could be provided the results for each utterance from both Google Translate and from our model and asked to rate both. A comparison of both the results would allow us to check the efficiency of our model more accurately.

3. Discussion:

Since there are no known existing models for translating Telugu-English mixed code to the best of my knowledge, we cannot compare this model to any existing models. Since a combination of Telugu-English text is usually used in informal settings and there are no known existing models to check the accuracy, the use of human participants might become more of a necessity rather than a choice. This model is expected to perform better than Google Translate as it uses a multi-step approach in order to reduce discrepancies by first translating the sentences and then correcting the grammar.

This model considers that all the words are written in English. Since Telugu words are written and spoken the same way it wouldn't cause much difference while transliteration. So, we can directly perform language identification on Telugu and English words and use them. Another possibility is when the Telugu words are written in Telugu and the English words are written in English. Another possibility which is not considered is when all the words are written in Telugu. In this case, the English words written in Telugu when transliterated back to English tend to cause errors as English words are not written in the same way as spoken. We would have to create a mapping of words and then use the most probable mapping which might give the right word in most cases. This model also doesn't deal with ambiguity when the words are annotated as both English and Telugu. In this case, we could translate the words into both languages and select the word which has a higher probability of being correct based on the sentence structure.

After analyzing the predicted translation, we have to incorporate different rules to choose which annotation to use. After collecting sufficient data, this model can also be extended to other Indian languages as most Indian languages are based on Sanskrit and like Telugu, have a sentence structure of SOV.

4. Word Count:

1590 words